UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE GRADO

ESTUDIO DE LA PARADOJA DE LA AMISTAD COMO HERRAMIENTA PARA EXPLORAR GRANDES REDES SOCIALES

Izaskun Dorronsoro Alberdi

Julio 2014

ESTUDIO DE LA PARADOJA DE LA AMISTAD COMO HERRAMIENTA PARA EXPLORAR GRANDES REDES SOCIALES

AUTOR: Izaskun Dorronsoro Alberdi

TUTOR: Manuel García-Herranz

Dpto. de Ingeniería Informática

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Julio 2014

Resumen

Desde hace unos años, el crecimiento del tamaño y la disponibilidad de la información en Internet han conducido a un nuevo tipo de estudio de fenómenos sociológicos basado en el análisis de grandes datos. Este tipo de análisis ha propiciado un nuevo auge de la teoría de redes, aumentando nuestra comprensión y capacidad de anticipación de fenómenos epidemiológicos de diferentes características. No obstante, el continuo crecimiento del tamaño de los datos en Internet, así como la creciente preocupación por factores relacionados con la privacidad anticipan un futuro en el que el análisis global de la información no será posible, siendo necesario encontrar mecanismos locales que permitan obtener conclusiones globales. La paradoja de la amistad (los amigos de una persona tienen, en media, más amigos que ella), enunciada por Feld, es un mecanismo local que ha abierto una vía de estudio mediante vistas localizadas permitiendo encontrar nodos con más grado que la media sin necesidad de conocer o analizar la red completa.

Este proyecto pretende profundizar en la paradoja de la amistad, analizando hasta dónde se puede penetrar en la distribución de grado utilizando dicha paradoja como mecanismo. También comprobaremos si penetra en las distribuciones de otras medidas de centralidad como k-coreness o betweenness y a qué velocidad lo hace. Y analizaremos las diferencias que existen en la obtención de grupos más centrales con distintas variantes de este mecanismo y un análisis global de la red.

Palabras clave

Paradoja de la amistad, medidas de centralidad, distribución de grado, betweenness, k-coreness, closeness.

Abstract

In the past years the growth and availability of digital data has led to a new way to study sociological phenomena using big data analysis. This kind of analysis has in turn caused a renewed interest on network theory to improve our understanding of epidemiological processes as well as our capacity to forecast them. However, the continuous increase in the size of digital data and the growing preoccupation about all things related to privacy may result in a future where global information analysis will not be possible and we will need to use local mechanisms to derive global conclusions. The Friendship Paradox, the fact that your friends have more friends than you do, first proposed by Feld, is a local procedure that has opened a way to analyze local views so that we can localize nodes that have a larger degree than the average without having to know or analyze an entire network.

We will study the Friendship Paradox in this work, studying how far we can analyze the degree distribution using the Paradox as a tool. We will also consider how it performs with other centrality measures such as k-coreness or betweenness and how fast it travels to more central nodes. We will also study whether there are differences when obtaining central groups with this mechanism and several variations of it and compare it with a global network analysis.

Index Terms

Friendship paradox, centrality measures, degree distribution, betweenness, k-coreness, closeness.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

INDICE DE ECUACIONES

1. Introducción

1.1. Motivación

Las redes sociales son comunidades de usuarios, que representan una estructura social mediante un grafo donde los vértices son las personas y las aristas son las relaciones (de amistad, laboral, otras) entre los individuos.

En el ámbito de Internet, son aplicaciones que permiten a sus usuarios estar en contacto con sus amigos o realizar nuevos amigos. Su fin es el poder compartir con los demás miembros información de diferentes temas, interactuar entre ellos y crear comunidades de personas con intereses comunes. Las redes sociales en Internet se pueden dividir en los siguientes tipos:

- **Redes sociales genéricas:** Son las redes más conocidas y más usadas. En ellas los usuarios pueden compartir contenidos de todo tipo. Ejemplos son Facebook, Twitter, o Tuenti.
- **Redes sociales profesionales:** Las relaciones que tienen los usuarios se basan en el ámbito laboral y pueden ser utilizadas para la búsqueda de empleo. Ejemplos: LinkedIn, Xing o Viadeo.
- **Redes sociales verticales o temáticas:** Estas redes se diferencian en que tienen una temática concreta y los miembros se suelen agrupar por hobbies similares, la búsqueda de pareja o las profesiones.

En los últimos años, las redes sociales online ha sufrido un aumento notable en número de usuarios y cantidad de información que por ellas se transmite. Se muestra a continuación la evolución en cifras de la red social Twitter.

Tabla 1: Número de usuarios de Twitter desde 2006 hasta 2012 [4]

	2006	2007	2008	2009	2010	2011	2012
Twitter	1.000	750.000	5.000.000	75.000.000	145.000.000	300.000.000	500.250.000

En este trabajo de fin de grado la red social seleccionada para realizar el análisis ha sido Twitter. Twitter [10] es un servicio de microblogging creado en marzo de 2006 por Jack Dorsey. Las interacciones entre los miembros de esta red se realizan mediante la publicación de *tweets*, mensajes de 140 caracteres en los cuales es posible compartir gran variedad de contenidos. Estos mensajes pueden estar etiquetados en temas y este etiquetado se realiza mediante *hashtags*, palabras precedidas por una almohadilla (#). Los usuarios pueden tener *followees* (gente a la que siguen para poder ver la información que cuelgan en la red) y *followers* (seguidores a los que se les muestra el contenido que el usuario publica, también denominados en algunos estudios como *amigos*).

Desde su nacimiento el crecimiento de Twitter ha ido en aumento, contando ahora con aproximadamente 500 millones de usuarios y generando 65 millones de tweets al día.

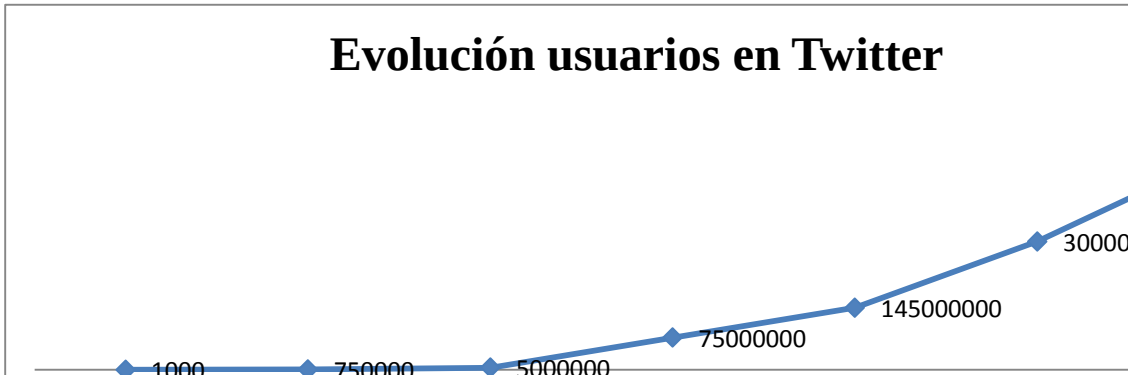


Ilustración 1. Evolución de los usuarios en Twitter desde 2006 hasta 2012 [4]

Como se puede observar en la Ilustración 1 el número de usuarios en Twitter ha aumentado en pocos años un 500249% (manteniendo un crecimiento de en torno al 100% en los últimos años). Esto ha conllevado que el poder recopilar toda la información de los miembros de la red sea una tarea complicada con un coste computacional, es decir, la

cantidad de recursos que deben ser utilizados para su análisis, inasumible en muchos casos. Dejando a un lado los costes computacionales o de transmisión de la información relacionados con el análisis de estos volúmenes de datos, la creciente preocupación por la privacidad de los usuarios dificulta y previsiblemente dificultará aún más en el futuro la obtención completa de la información y estructura de la red para su posterioranálisis [2].

Por ello la búsqueda de mecanismos locales que permitan obtener conclusiones globales sin necesidad de disponer ni analizar la información al completo se ha convertido en una prioridad.

1.2. Alcance y objetivos

La paradoja de la amistad (los amigos tienen en media más amigos que uno) es una curiosidad matemática planteada por Scott Feld [12] que ha sido utilizada recientemente con prometedores resultados con el fin de realizar análisis globales a partir de información local [11]. No obstante sus potencialidades y limitaciones no han sido aún analizadas en profundidad.

Así, utilizando la paradoja de la amistad como herramienta, en este trabajo de fin de grado se pretende profundizar en esta teoría estudiando una muestra de datos reales obtenidos de la red social Twitter para poder resolver las siguientes preguntas:

1. ¿Tienen también los amigos de mis amigos más amigos que mis amigos? En tal caso, se estudiará hasta dónde se puede penetrar en la distribución de grado aplicando la paradoja de la amistad de forma recursiva.

2. ¿Penetra la paradoja de la amistad sólo en la distribución de grado o lo hace también en las distribuciones de otras medidas de centralidad como k-coreness, closeness o betweenness? ¿Lo hace con la misma velocidad?

3. ¿Qué repercusión tiene la forma de elegir los amigos? ¿En coste? ¿En adquisición de centralidad?

4. ¿Qué diferencias existen entre la obtención de grupos más centrales con este mecanismo y mediante un análisis global de la red? Se estudiará qué coste tiene obtener un grupo de usuarios más central utilizando la paradoja de la amistad y sin utilizarla, para poder comparar el coste de cada opción y sacar conclusiones de cuál de los dos métodos es más eficiente a la hora de encontrar a los miembros centrales de una red grande.

1.3. Estructura del documento

En el capítulo 1, Introducción, se ha realizado una breve introducción a las redes sociales, en especial a Twitter, se da la motivación para este trabajo de fin de grado y el alcance y objetivos que se van a resolver al finalizar esta memoria.

Más adelante en el capítulo 2, Estado del arte, se explica qué es la paradoja de la amistad, en qué estudios ya realizados se ha utilizado y cuáles son sus ventajas.
Después se procederá a explicar el diseño y el desarrollo de cómo se va a realizar el análisis sobre Twitter, todo ello recogido en el capítulo 3, Diseño y desarrollo, exponiendo seguidamente los resultados obtenidos al llevar a cabo las pruebas diseñadas en el capítulo 4, Pruebas y resultados.

Por último, en el capítulo 5, Conclusiones y trabajo futuro se recogen las conclusiones de la aplicación de la paradoja de la amistad como herramienta en el estudio de grandes redes sociales y el trabajo futuro que se podría llevar a cabo.

2. Estado del arte

2.1. Paradoja de la amistad

La paradoja de la amistad es un fenómeno que fue observado en 1991 por el sociólogo Scott L. Feld[12] que determina que, en media, tus amigos tienen más amigos que tú. Lo explicó mediante un ejemplo muy sencillo en el que hay ocho niñas y sus correspondientes amigos:

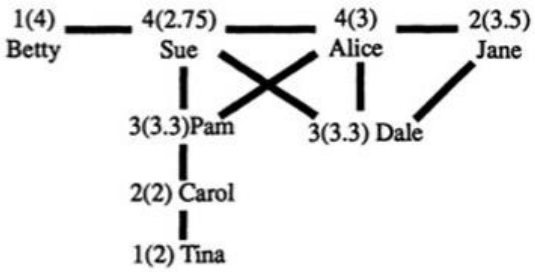


Ilustración 2. Ejemplo paradoja de la amistad (ilustración obtenida del artículo de Feld[12])

	Number of Friends (x_i)	Total Number of Friends of Her Friends (Σx_j)	Mean Number of Friends of Her Friends ($\Sigma x_j/x_i$)
Betty	1	4	4
Sue	4	11	2.75
Alice	4	12	3
Jane	2	7	3.5
Pam	3	10	3.3
Dale	3	10	3.3
Carol	2	4	2
Tina	1	2	2
Total	20	60	23.92
Mean	2.5*	3†	2.99*

* For eight girls.
† For 20 friends.

Tabla 2: Resumen del número de amigos y de la media de los amigos de los amigos para cada una de las niñas de la Ilustración 2

El número que aparece junto a los nombres en la Ilustración 2 indica cuantos amigos tiene cada niña y el número entre paréntesis es la media de amigos que tienen sus amigos.
El número medio de amigos que tienen las ocho niñas es de 2.5 amigos, en cambio al calcular la media de los amigos que tienen sus amigos se obtiene 3 como se puede observar en la Tabla 2.

Una red social está representada por un grafo no dirigido *G = (V,E)* donde el conjunto de vértices V corresponde a la gente de la red social y el conjunto de aristas E a la relación de amistad entre pares de personas.

Analizando el estudio de Feld se observa que se trata de una anécdota matemática que se cumple para toda la red, si el usuario *k* tiene *n* amigos entonces *k* se contará *n* veces al realizar el cálculo de sus amigos. Si se mide matemáticamente la media de conexiones por vértice en el grafo G [12] se obtiene la siguiente ecuación:

$$\mu = \frac{\sum_{v \in V} k_v}{|V|} = \frac{2|E|}{|V|}$$

Ecuación 1. Media de conexiones por vértice

Para calcular la media de los amigos de los amigos [12] (conexiones de grado dos) será:

$$\rho = \frac{\sum_{v \in V} k_v^2}{\sum_{v \in V} k_v}$$

Ecuación 2. Media de conexiones grado 2

Con estas dos ecuaciones se concluye que la distribución de grado para conexiones de grado 2 esperada será la función de distribución de grado de un vértice *P(k)* multiplicado por las veces que *k* aparecerá en el número total de amigos de sus amigos, es decir *Q(k)=kP(k)* donde *Q(k)* es la distribución de grado de los amigos y *P(k)* la de la red.

2.2. Análisis de la paradoja de la amistad para una muestra de la red completa

No obstante para ciertas aplicaciones este método pudiera no resultar del todo adecuado, ya que se podría interpretar que no es que los amigos tengan más grado sino que se están contando varias veces a los indivíduos con mayor grado. Así, surge una nueva pregunta: si se eliminasen los duplicados, ¿el número medio de amigos de mis amigos sigue siendo mayor? La respuesta es no si se coge toda la red ya que el conjunto de amigos de una red es, aproximadamente, el conjunto de los nodos de la red, pero si se coge una porción gamma de la red sí que funciona con y sin repetidos, como se puede observar en *Using Friends as Sensors to Detect Global-Scale Contagious Outbreaks*[11] donde se demostró que la paradoja de la amistad sigue funcionando cuando no se cuentan usuarios repetidos (ver Ilustración 3), posibilitando la obtención de resultados globales haciendo un análisis local.

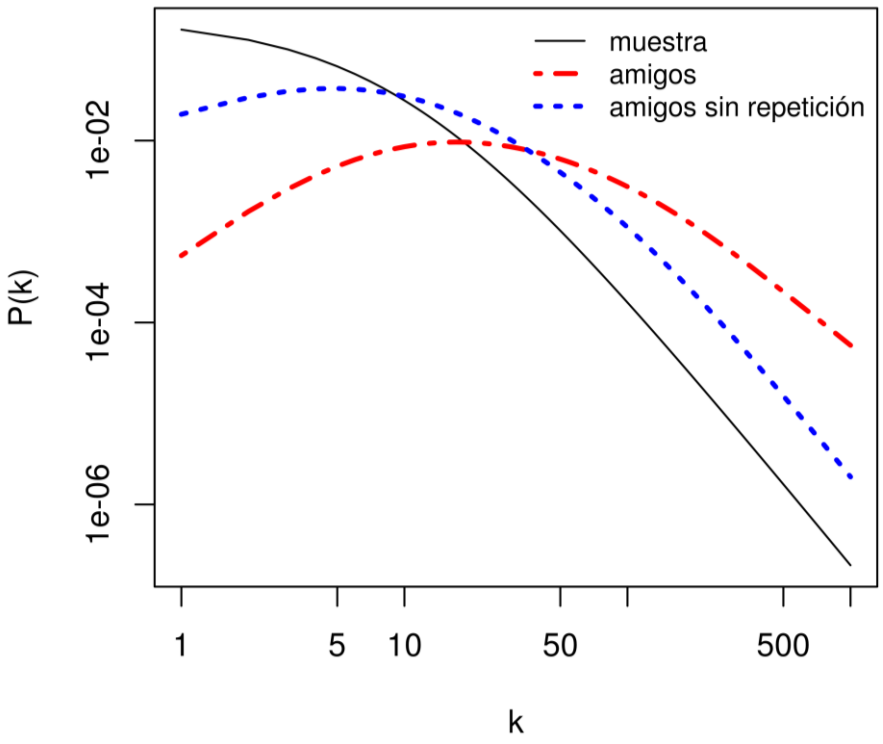


Ilustración 3. Distribución de grado teórica [11]

2.3. Aplicaciones

El análisis de la paradoja de la amistad implica que al estudiar una muestra de una red los amigos obtenidos de dicha muestra tienden a ser más centrales en cualquier tipo de red en la que exista una varianza en el grado de sus nodos. Esta observación se ha utilizado como una manera de predecir y retrasar el curso de las epidemias [1] mediante el uso de este proceso de selección aleatorio para elegir a los individuos a inmunizar o monitorizar la infección evitando al mismo tiempo la necesidad de realizar un cálculo complejo de la centralidad de todos los nodos de la red.

James Fowler y Nicholas Chritakis, en un estudio realizado en el campus de Harvard [9] encontraron la manera de detectar brotes de gripe casi 2 semanas antes de que las medidas tradicionales de vigilancia lo hicieran. Cogieron un grupo de estudiantes (grupo control) y un subgrupo de sus amigos (grupo sensor). Al realizar un seguimiento de cómo afectaba la gripe para cada uno de los grupos observaron que el grupo sensor estaba expuesto a ser infectado antes que el grupo control.

Los nodos centrales de una red están también expuestos a recibir más información que el resto y, correlacionando con el grado, a transmitirla antes. Usando éste último fenómeno se realizó un estudio sobre más de 50 millones mensajes de Twitter publicados antes, durante y después del huracán Sandy [8], que afecto a varios países americanos, entre ellos, Colombia, Venezuela, Estados Unidos y Canadá, que duro del 22 al 29 de octubre de 2012 y en el que murieron 287 personas. Siguiendo la paradoja de la amistad se seleccionaron dos grupos, sensor y control, y se observó que la diferencia de centralidad entre usuarios se traduce en una ventaja de conocimiento moderada (de hasta de 26 horas) incluso en eventos tan exógenos como un huracán.

En la línea de este trabajo se han realizado estudios preliminares de en qué otras características se observa esta paradoja, observando, en palabras de Hodas [3] que tus amigos son más interesantes que tú.

3. Diseño y desarrollo

El diseño y desarrollo de las pruebas se realiza sobre una base de datos que se obtuvo al recopilar los datos de Twitter entre el Junio y Diciembre de 2009, tanto de la red de seguidores y seguidos como de sus tweets [13]. La base de datos analizada cuenta con 41,7 millones de usuarios, cerca de 1.470 millones de relaciones entre ellos y 106 millones de tweets.

Para poder realizar los análisis sobre muestras de la red se van a crear tres scripts, que se van a codificar en AWK [5] ya que es un lenguaje diseñado para procesar datos basados en texto línea a línea, permitiendo el análisis de grandes ficheros de datos que de otra forma son intratables. Además, se ha precompilado la información de grado de la red un un solo fichero (users.degrees) con una línea por usuario indicando su número de seguidores (in-degree) y de seguidos (out-degree). Los scripts serán los siguientes:

- getRnd.awk: para extraer una muestra aleatoria, sin repetición, del tamaño que se desee, de una lista de identificadores.
- getSampleDegree.awk: devuelve la porción del fichero usersdegree correspondiente a una lista de usuarios especificada por parámetros.
- getFriends.awk: retorna una lista con todos los seguidos (con repetidos) por una determinada lista de usuarios.

3.1. La paradoja de la amistad

Existen muchas maneras de seleccionar los amigos de una muestra y dependiendo de cómo se realice esta operación cabe esperar que los resultados varíen. Por eso para la obtención de los amigos se van a realizar tres estrategias diferentes que producen diferentes muestras:

1. Replicando los análisis de [11] y buscando comprender cómo estos se extienden a otras medidas de centralidad y a sucesivas iteraciones de la paradoja el primer método consiste en seleccionar un grupo aleatorio de entre la lista completa de amigos sin repetición.
2. Buscando comparar los resultados de [11] con los de [12] el segundo método consiste en elegir una muestra aleatoria (en la que puede haber repetidos) de entre la lista completa de amigos con repetición.
3. Para explorar una vía que extraiga las ventajas de [12] sin que la muestra final pueda mostrar grados medios artificiales (debidos a que se está computando el grado de un solo indivíduo varias veces) el tercer método consistirá en extraer una muestra aleatoria sin repetidos de entre la lista completa de amigos con repetidos.

Para ver cómo afecta la paradoja de la amistad según se va penetrando en la red, o sea, se van a estudiando iteraciones sucesivas de grupos de amigos, se van a diseñar diferentes análisis:

1. Comparativa de la distribución de grado de una muestra de todos los amigos del mismo tamaño que la muestra inicial.

Para el análisis de la evolución de la paradoja de la amistad sobre diferentes iteraciones se realizará el siguiente muestreo. Sobre una muestra de tamaño 50.000 de la red completa (de 40 millones de usuarios aprox.), a la que denominaremos *Muestra* se van a obtener tres grupos:

- *Amigos*: es un grupo formado por una muestra del mismo tamaño que la inicial realizada sobre todos los amigos de *Muestra.*
- *Amigos*^{2}: se seguirá el mismo procedimiento que el utilizado para el grupo anterior, pero ahora los amigos se obtendrán de los usuarios que integran *Amigos.*
- *Amigos*^{3}: ahora la muestra a seleccionar se recoge de *Amigos*^{2}.

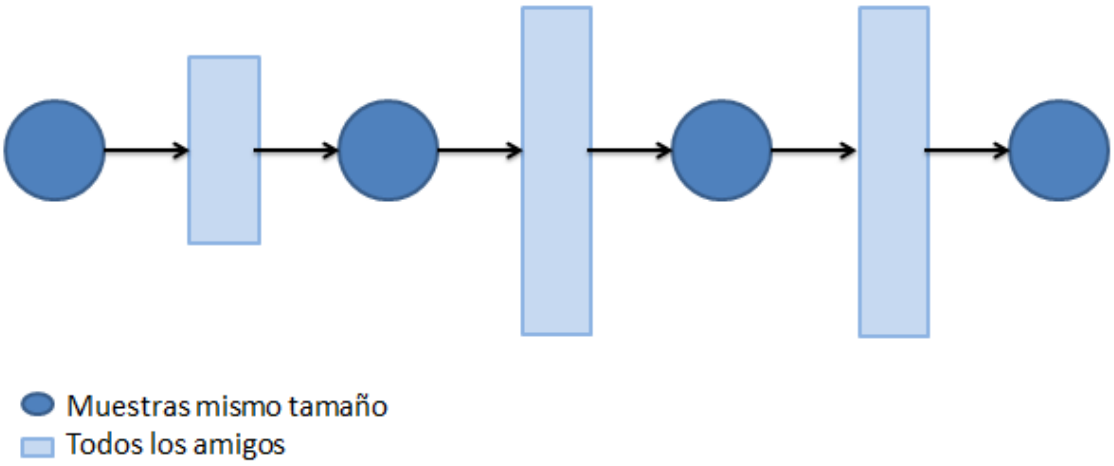


Ilustración 4. Esquema para la comparación del grado medio

2. Comparación del grado medio de 100 muestras.

Siguiendo el esquema anterior se van a obtener los cuatro grupos *Muestra*, *Amigos*, *Amigos*^{2} y *Amigos*^{3} pero ahora para 100 muestras de un tamaño de 50.000 usuarios elegidos aleatoriamente y obteniendo tanto amigos con repetición como sin ella.

3.2. Otras medidas de centralidad

Una vez analizada la paradoja de la amistad en cuanto a grado se refiere, se realiza un análisis para poder obtener información sobre otras medidas de centralidad y así observar si se penetra más con estas otras medidas.

El cálculo de muchas medidas de centralidad como la betweenness o el k-coreness es computacionalmente muy pesado por lo que es necesario reducir el tamaño de la red a analizar. Por ello ahora se seleccionan redes de usuarios que han mencionado alguno de los *hashtags*: **#09memories, #fun140, #iwillneverforget** y **#wheniwaslittle**. Estos *hashtags*

han sido seleccionados de entre los más usados en el segundo semestre de 2009 que, además no han sido usados en el mes de julio (asegurándonos así de que tenemos registrado su nacimiento en Twitter) [13]

Para realizar el cálculo de otras medidas de centralidad se va a hacer uso de la biblioteca *igraph*, incluida en R [6], la cual posee las funciones necesarias para poder medir la betweenness, el k-coreness y el closeness.

3.2.1. Betweenness

Con la medida de centralidad betweenness se mide cuántos caminos mínimos pasan por un determinado nodo. Esta medida de centralidad viene a determinar la importancia de un nodo en la transmisión de la información por la red y, por tanto, cómo se verían afectadas las comunicaciones entre el resto de los nodos en caso de que este desapareciera.

3.2.2. K-Coreness

Con K-Coreness se mide cómo de importantes son las personas a las que un nodo está conectado, es decir, cuánto grado tienen las personas conectadas a él. El k-core de un grafo es un subgrafo máximo en el que cada vértice tiene al menos un grado k. El coreness de un vértice es k si pertenece a la k-core, pero no a la (k +1)-core.

3.2.3. Closeness

Closeness mide cuántos pasos son necesarios para acceder a todos los demás vértices desde un vértice dado. La proximidad central de un vértice se define por la inversa de la longitud media de los caminos más cortos a todos los otros vértices en el grafo.

$$\frac{1}{\sum dist(v,i)_{v!=i}}$$

Ecuación 3. Calculo de closeness

Utilizando las funciones betweenness, k-coreness y closeness se elaborarán gráficas sobre diferentes redes de *hashtags* aplicándoles las diferentes estrategias en la obtención de los amigos, con las que poder comprobar si la paradoja de la amistad penetra también en estas medidas de centralidad y se comparará si la velocidad de penetración es la misma que con la distribución de grado.

3.3. Análisis de costes

Se quiere comparar qué diferencias de costes existen entre realizar un análisis global de la red o aplicar la paradoja de la amistad para la obtención de grupos de usuarios más centrales, centrándose el estudio en conseguir una distribución de grado similar.

Primero se obtendrá un grupo de un tamaño de 50.000 usuarios de los cuales se obtendrán sus amigos y sobre estos se escogerá aleatoriamente una muestra del mismo tamaño que la inicial (*rnd_friends*) y otra muestra con los 50.000 amigos de mayor grado (*top_friends*). Más tarde, cogiendo diferentes tamaños de muestra se intentará conseguir un grupo de usuarios con una distribución de grado similar a *top_friends* y será entonces cuando se podrá observar si existen diferencias a la hora de obtener grupos más centrales. Los tamaños de las muestras se escogerán aleatoriamente.

4. Pruebas y resultados

4.1. Análisis de la paradoja de la amistad

La paradoja de la amistad afirma que para una red completa el grupo de amigos en media tienen más amigos que los nodos de la red. Más tarde se demostró en *Using Friends as Sensors to Detect Global-Scale Contagious Outbreaks* [11] que la paradoja se sigue cumpliendo para una muestra de la red. Es por eso que lo primero a realizar va a ser replicar el estudio de si para una muestra de la red y sus amigos, estos últimos son más centrales en la red comparando la distribución de grado de una muestra aleatoria de la red con una de igual tamaño de sus amigos.

En la Ilustración 5 se puede observar cómo se cumple la paradoja, observándose que la distribución de grado de la muestra de amigos presenta una menor cantidad de usuarios con grado bajo, en comparación con la muestra inicial, a favor de una mayor densidad de usuarios con grados más altos, tal y como predecía el modelo teórico (ver Ilustración 3).

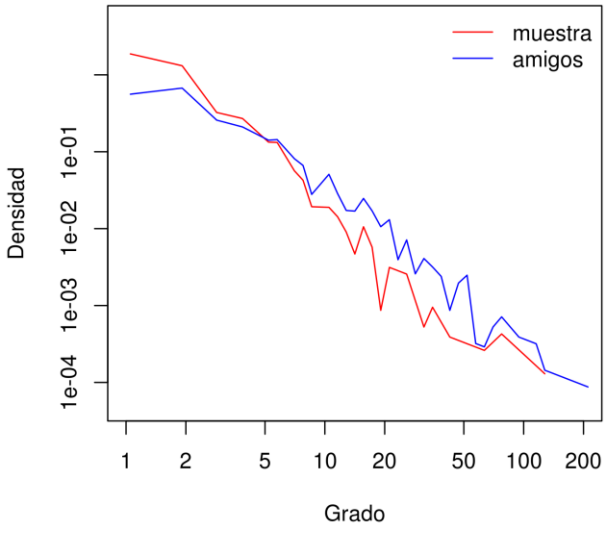


Ilustración 5. Distribución de grado de una muestra seleccionada aleatoriamente de una red y de sus amigos.

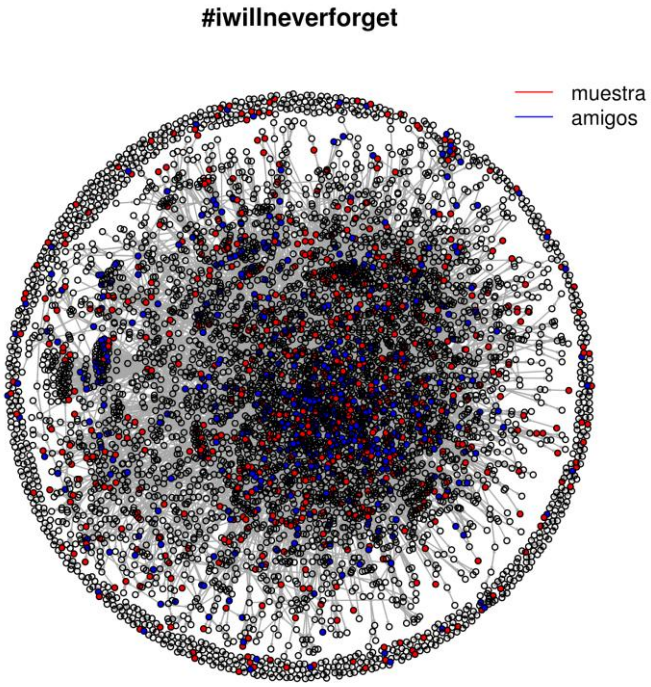


Ilustración 6. Red #iwillneverforget mostrando la muestra (rojo) y sus amigos (azul)

Gráficamente, se puede observar en la Ilustración 6 cómo el grupo de amigos (puntos azules) tiende a localizarse en lugares más centrales de la red, mientras que los usuarios de la muestra inicial (puntos rojos) se encuentran más dispersos.

Una vez observado que la paradoja de la amistad permite obtener un grupo con mayor grado se analizará a continuación si la paradoja de la amistad permite seguir adquiriendo centralidad en sucesivas iteraciones. Así, para una muestra aleatoria de usuarios de la red completa de Twitter, sus amigos (*amigos*), los amigos de sus amigos (*amigos*^{2}) y los amigos de los amigos de sus amigos (*amigos*^{3}), obtenidos todos ellos de forma aleatoria de entre el conjunto completo de amigos sin repetición, se estudia si la paradoja de la amistad penetra en la distribución de grado. En la Ilustración 7 se puede observar que la distribución de grado en *amigos*^{2} y *amigos*^{3} es muy similar a la del grupo *amigos* aunque no igual, como se ve en la Ilustración 8 que al exponer el grado medio de 100 muestras distintas obtenidas sin repetición para los grupos *muestras, amigos, amigos*^{2} *y amigos*^{3} se observa que el grado a partir del grupo *amigos* decrece.

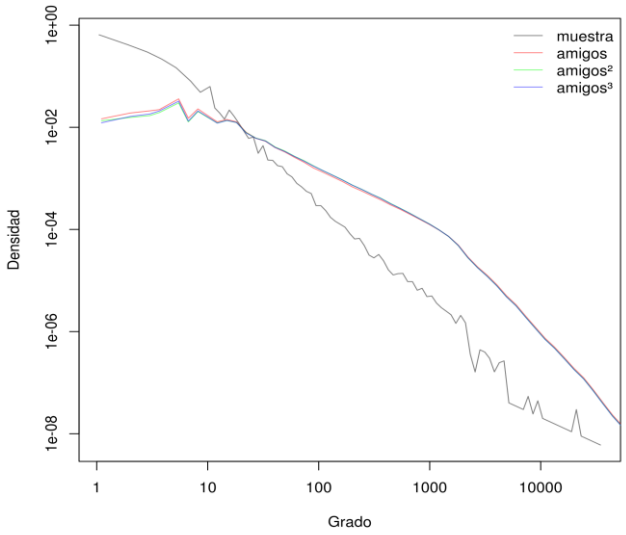


Ilustración 7. Comparación de la distribución de grado para una muestra aleatoria de usuarios, sus amigos, los amigos de sus amigos (*amigos*^{2}) y los amigos de los amigos de sus amigos (*amigos*^{3}) obteniendo los amigos sin repetición

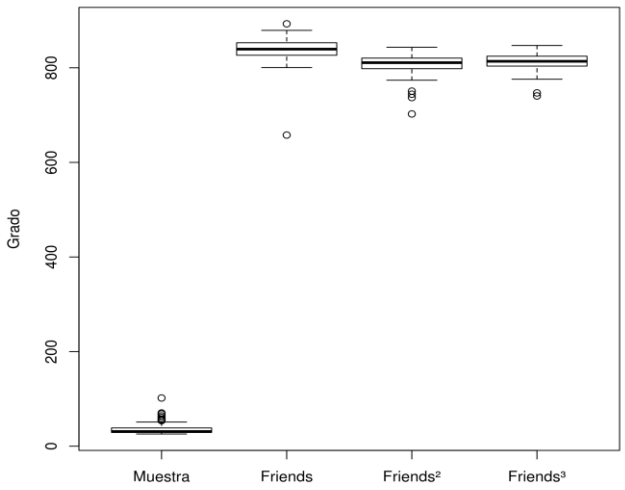


Ilustración 8. Grado medio calculado sobre 100 muestras aleatorias, sus amigos, los amigos de sus amigos (*amigos*^{2}) y los amigos de los amigos de sus amigos (*amigos*^{3}) obteniendo los amigos sin repetición

Así, podemos ver cómo la ganancia de centralidad se produce en la primera iteración de la paradoja de la amistad, no en las siguientes en las que puede observarse incluso una ligera caída.

4.2. Otras medidas de centralidad

Aunque se ha demostrado que para las distribuciones de grado la paradoja de la amistad funciona, resulta interesante ver si también se cumple en el cálculo de otras medidas de centralidad. Como ya se expuso en el apartado 3.2. Otras medidas de centralidad, necesitamos reducir el tamaño de la red a explorar para poder realizar el cálculo de diversas medidas de centralidad. Por ello analizamos la betweenness, el k-coreness y el closeness en las siguientes redes de *hashtags* (la subred de Twitter en la que todos los nodos han usado en algún momento un determinado *hashtag*).

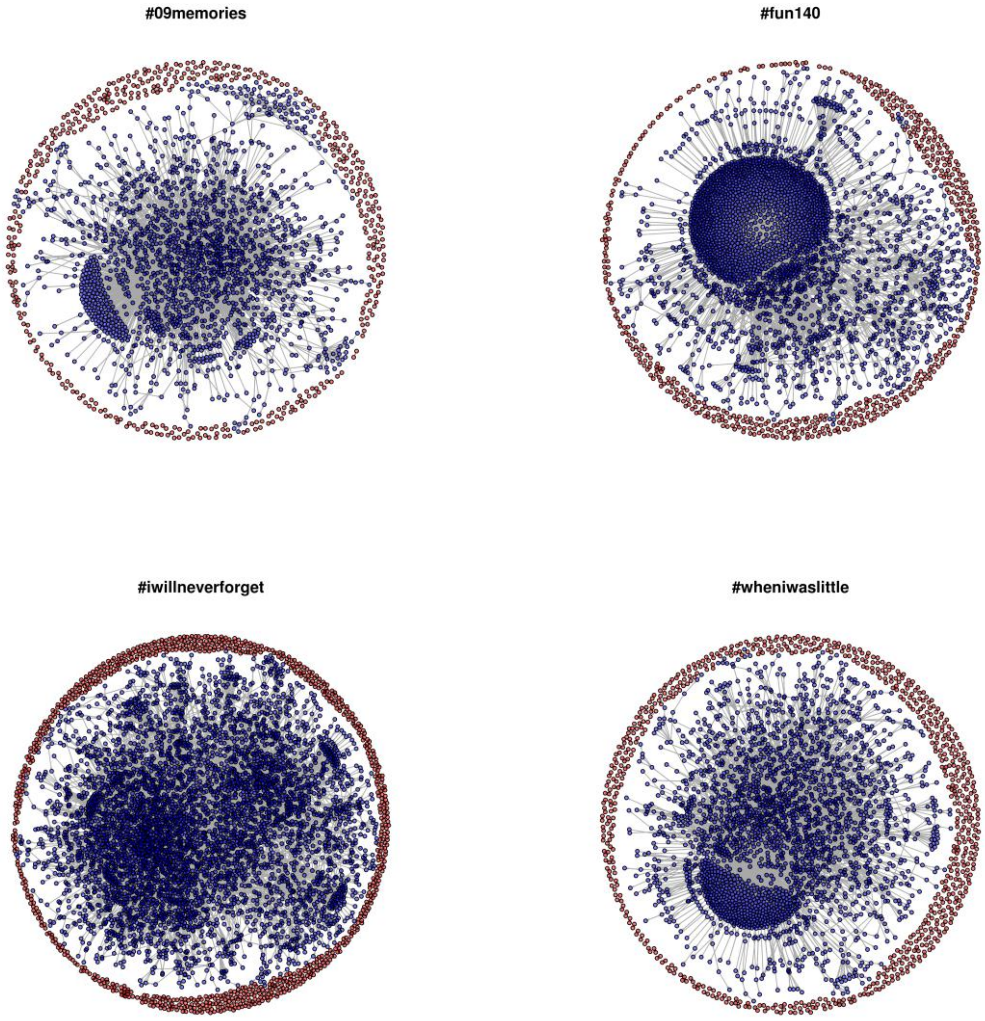


Ilustración 9. Redes de usuarios que han utilizado el mismo hashtag. Se muestra en azul la comunidad central de la red y en rojo los nodos que no están conectados a ella.

Antes de estudiar el resto de medidas de centralidad, se examina la distribución de grado para la red *#iwillneverforget* para asegurarnos que los resultados obtenidos en el apartado anterior para la red completa de Twitter siguen cumpliéndose para estas subredes. La estrategia para la obtención de amigos que se va a utilizar será la misma usada para toda la red anteriormente, que descarta duplicados. Como se puede ver en la Ilustración 10 y la Ilustración 11 los resultados son los mismos que para la red completa de Twitter, habiendo una gran diferencia entre el grado de la muestra y la de sus *amigos*, no observándose

crecimiento en sucesivas iteraciones. Así, se sigue cumpliendo la paradoja de la amistad pero para más profundidad de amigos la distribución de grado se mantiene o incluso cae ligeramente.

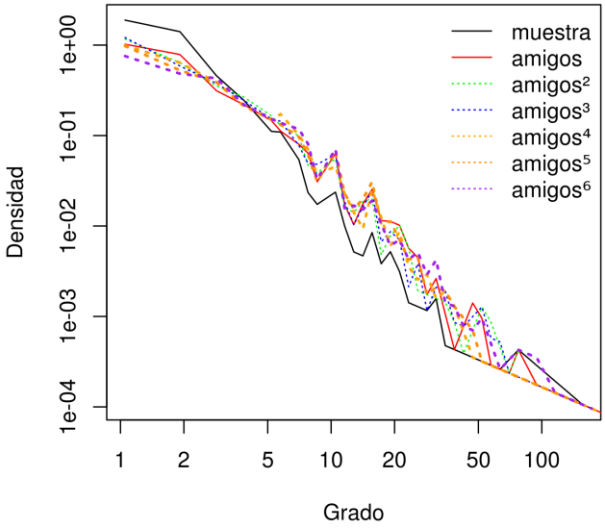


Ilustración 10. Comparación de la distribución de grado para una muestra aleatoria de usuarios, *amigos*, *amigos*^{2}, *amigos*^{3}, *amigos*^{4}, *amigos*^{5} y *amigos*^{6} habiendo obteniendo los amigos sin repetición en la red del hashtag #iwillneverforget

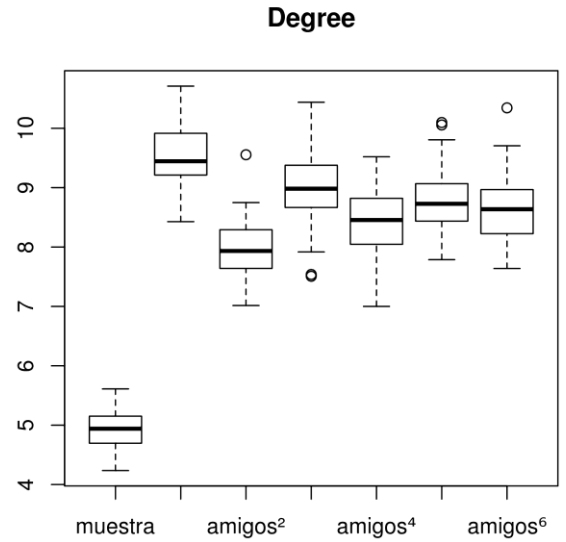
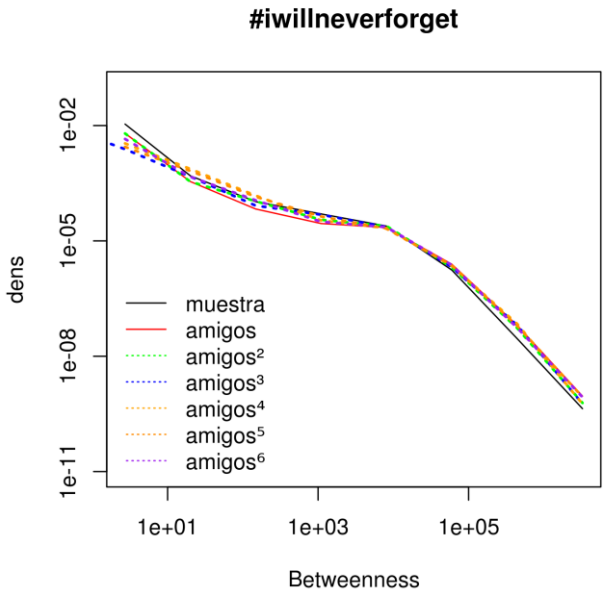
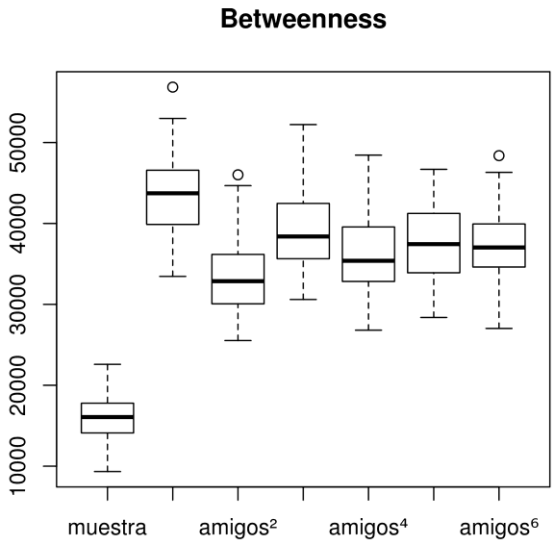
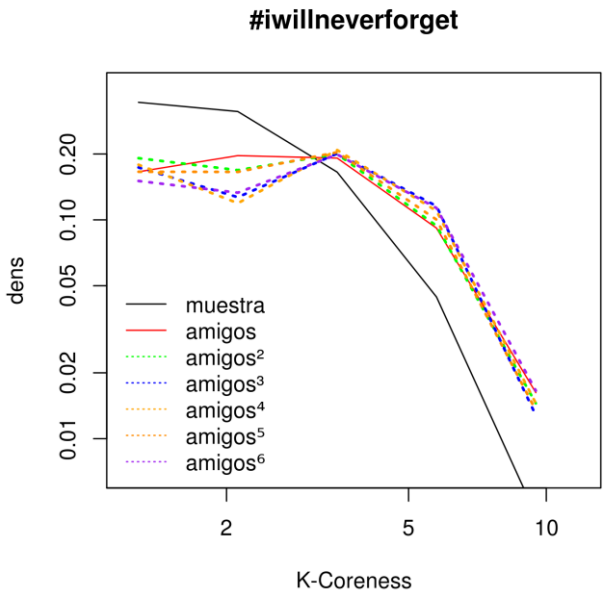


Ilustración 11. Grado medio calculado sobre 100 muestras aleatorias de *muestra*, *amigos*, *amigos*^{2}, *amigos*^{3}, *amigos*^{4}, *amigos*^{5} y *amigos*^{6} habiendo obteniendo los amigos sin repetición en la red del hashtag #iwillneverforget

Tanto en el análisis de la betweenness como en el k-coreness pasa igual que con el grado, el grupo *amigos* penetra y posee una gran diferencia del grupo *muestra*, pero a partir de *amigos* ambas medidas decrecen o se mantienen constantes para los grupos siguientes (*amigos*^{2}, *amigos*^{3}, *amigos*^{4}, *amigos*^{5} y *amigos*^{6})., como se puede ver en las Ilustraciones 12 y 13.







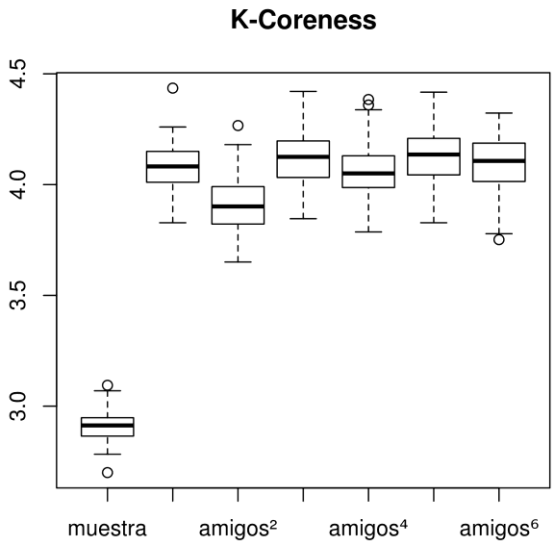


Ilustración 12. Distribución de bewteennness y k-coreness de una muestra aleatoria, *amigos*, *amigos*^{2}, *amigos*^{3}, *amigos*^{4}, *amigos*^{5} y *amigos*^{6} para amigos obtenidos sin repetición.

Ilustración 13. Bewteennness y k-coreness medio obtenido de 100 muestras aleatorias de *muestra*, *amigos*, *amigos*^{2}, *amigos*^{3}, *amigos*^{4}, *amigos*^{5} y *amigos*^{6} para la estrategia de obtención de amigos sin repetición.

El closeness, en cambio, presenta un comportamiento diferente, con un salto a partir de *amigos*^{2} manteniendo un crecimiento asintótico que se mantiene a partir de *amigos*^{3}. Aunque no exista evidencia estadística para confirmarlo, se asemeja bastante a una distribución logarítmica.

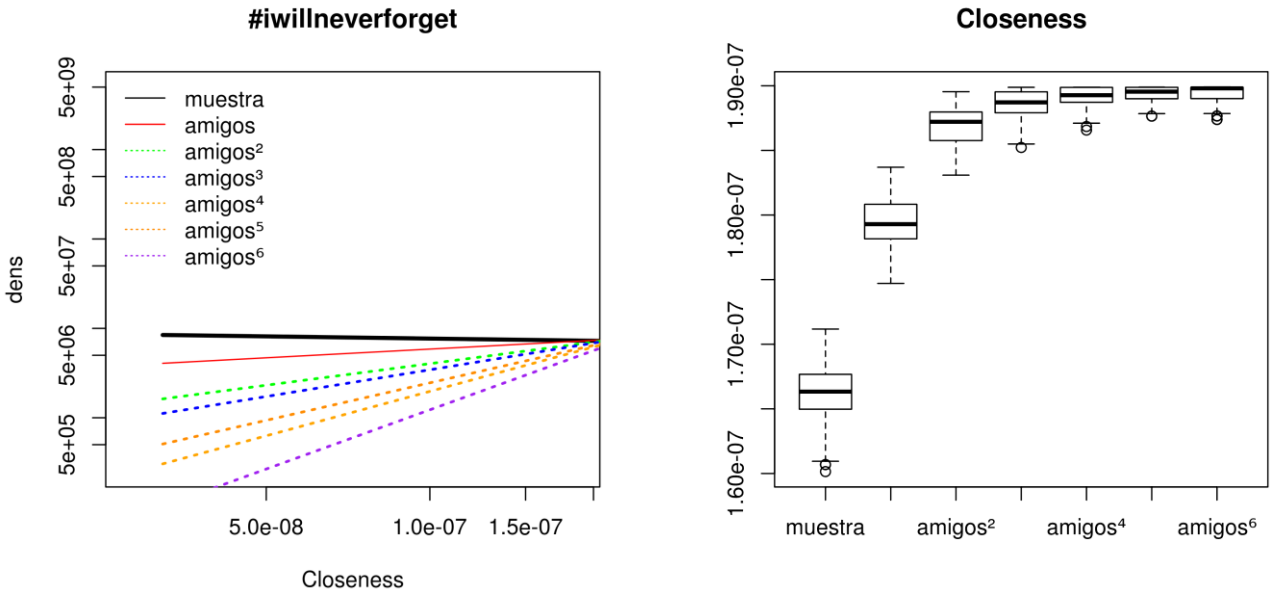


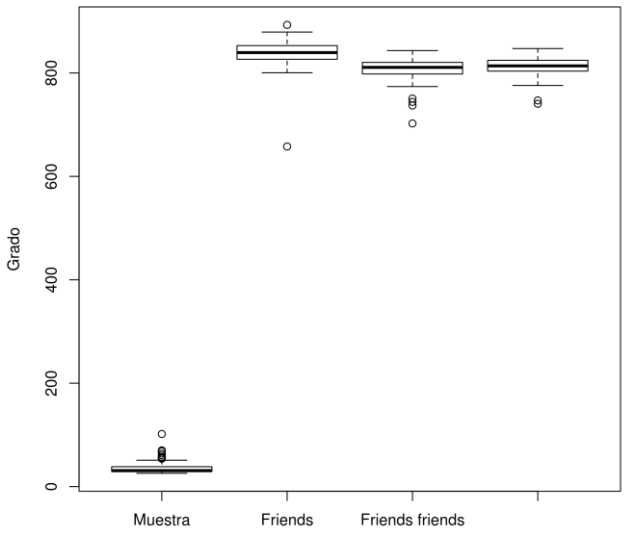
Ilustración 14. Distribución de Closeness para una muestra y boxplots para mostrar la media de 100 muestras aleatorias para *muestra, amigos, amigos*^{2}, *amigos*^{3}, *amigos*^{4}, *amigos*^{5} y *amigos*^{6} para amigos obtenidos sin repetición.

4.3. Variantes de la paradoja de la amistad

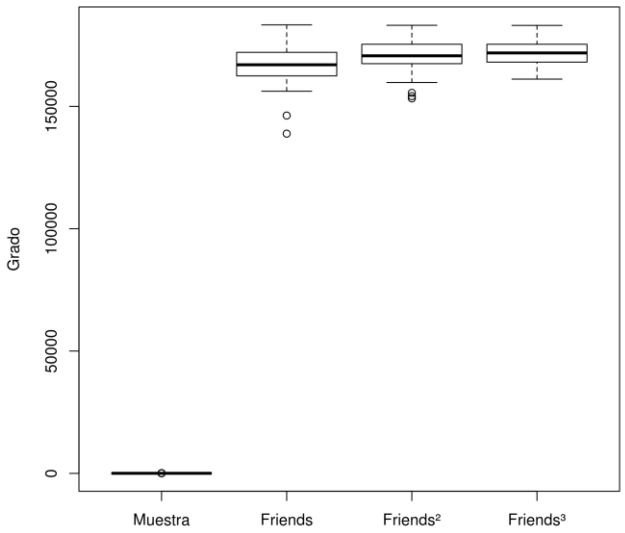
Como existen varias formas de seleccionar los amigos, cabría esperar que estos cambios puedan hacer que los resultados varíen. Por ello se analiza la paradoja de la amistad cambiando la estrategia de obtención de amigos de las muestras aleatorias para considerar, como en la paradoja original de Feld, los amigos con duplicados.

Comparando la media de grado sacada de 100 muestras aleatorias para amigos con repetición y sin repetición se observa (ver Ilustración 15) que el grado medio para los amigos duplicados es mucho más alto, superando el grado 150.000, que para los sin repetidos que tienen como grado 800.

Aparte cuando se escogen amigos con repetición para los grupos *amigos*^{2} y *amigos*^{3} la distribución de grado se mantiene o incluso crece ligeramente. Cuando no se escogen duplicados a partir de *amigos*^{2} el grado medio suele tener una tendencia ligeramente decreciente.



Amigos sin repetición



Amigos con repetición

Ilustración 15. Comparación del grado medio para 100 muestras con y sin repetición

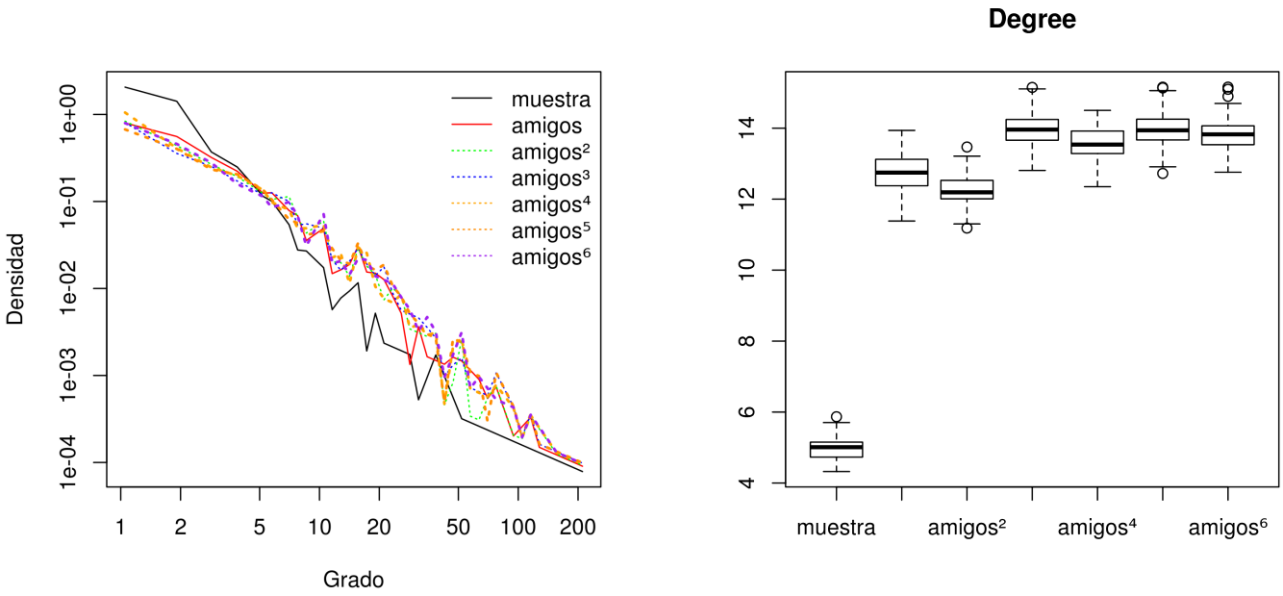


Ilustración 16. Distribución de grado y grado medio de 100 muestras obteniendo los amigos con repetidos

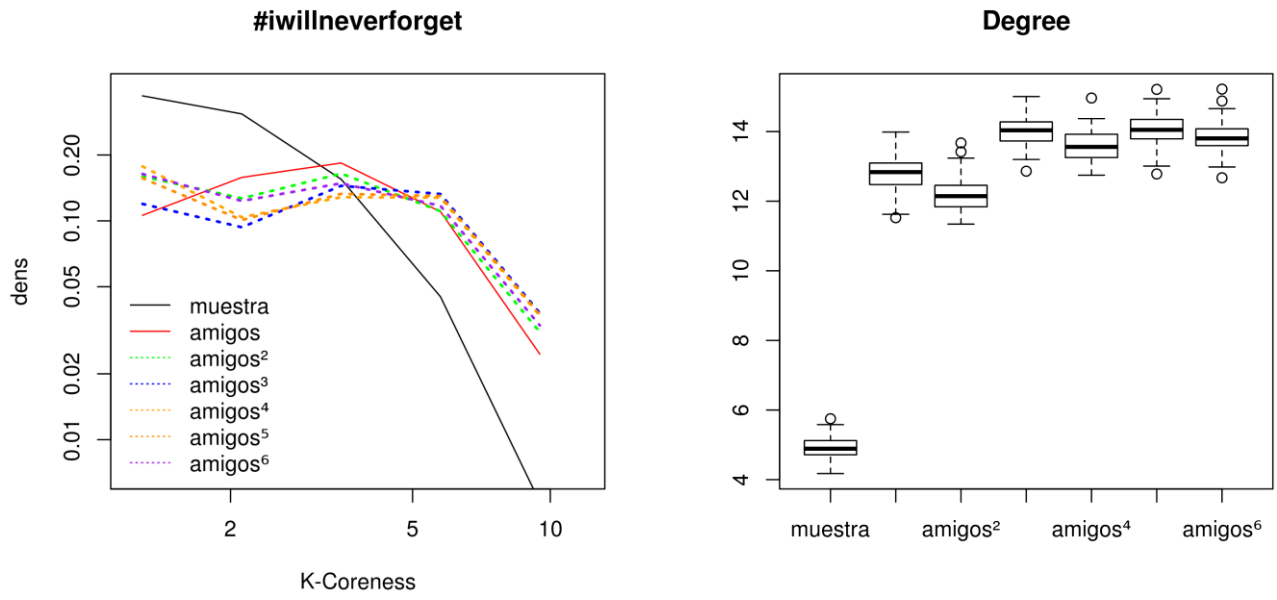


Ilustración 17. Distribución de grado y grado medio de 100 muestras en las cuales se obtienen todos los amigos y después se eliminan los duplicados.

Eliminando los repetidos se sigue obteniendo un grado alto aunque menor que si se escogen duplicados. Esto ocurre porque en la muestra con repetición las personas que tienen más probabilidad de aparecer son los de grado más alto. Aunque sin eliminar repetidos se obtenga un grado alto, esta centralidad es artificial ya que están apareciendo muchas veces gente con mucho grado.

4.3.1 Otras medidas de centralidad

Se va a volver a realizar el estudio de otras medidas de centralidad para poder analizar cómo se comportan dichas medidas cuando varían las estrategias de obtención de amigos.

4.3.1.1 Betweenness

Tanto si se utiliza la estrategia de coger amigos con repetición y luego obtener una muestra sin repetición, o por el contrario se deja la muestra con repetidos, el grupo *amigos* y $amigos^2$ son muy parecidos en cuanto a la betweenness media y se observa una tendencia creciente con el grupo $amigos^3$ el cual será el grupo para el cual la betweenness penetra más en las dos estrategias de obtención de amigos. A partir de $amigos^3$ se mantiene la misma betweenness para los grupos que siguen, o sea, $amigos^4$, $amigos^5$ y $amigos^6$.

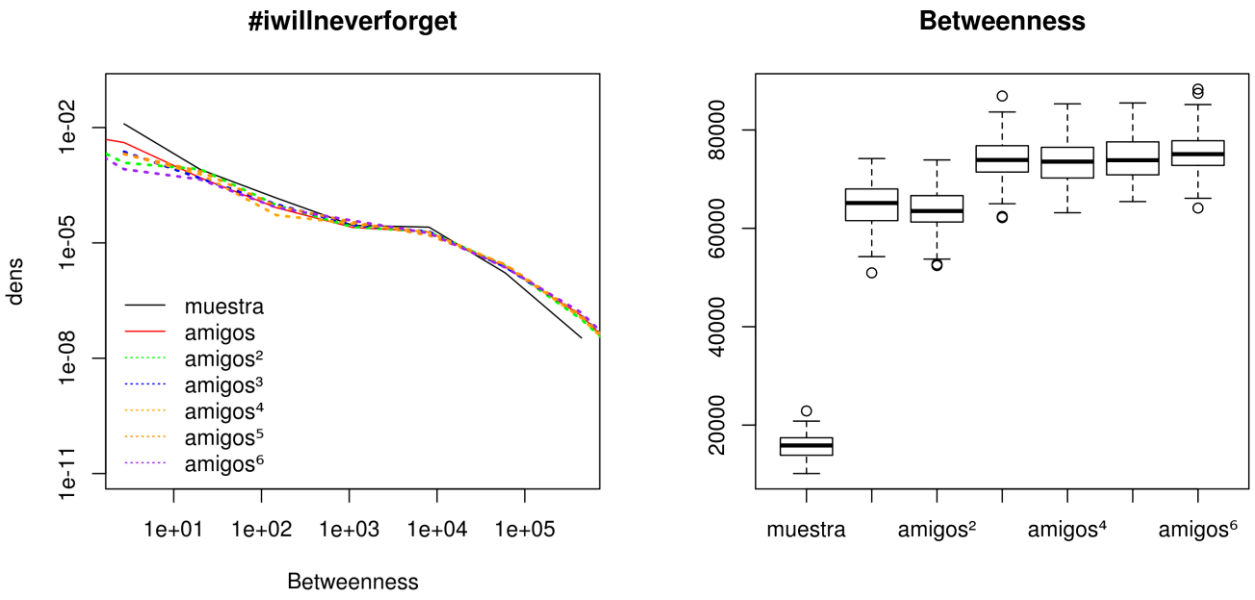


Ilustración 18. Distribución de la betweenness y betweenness medio para amigos con repetición

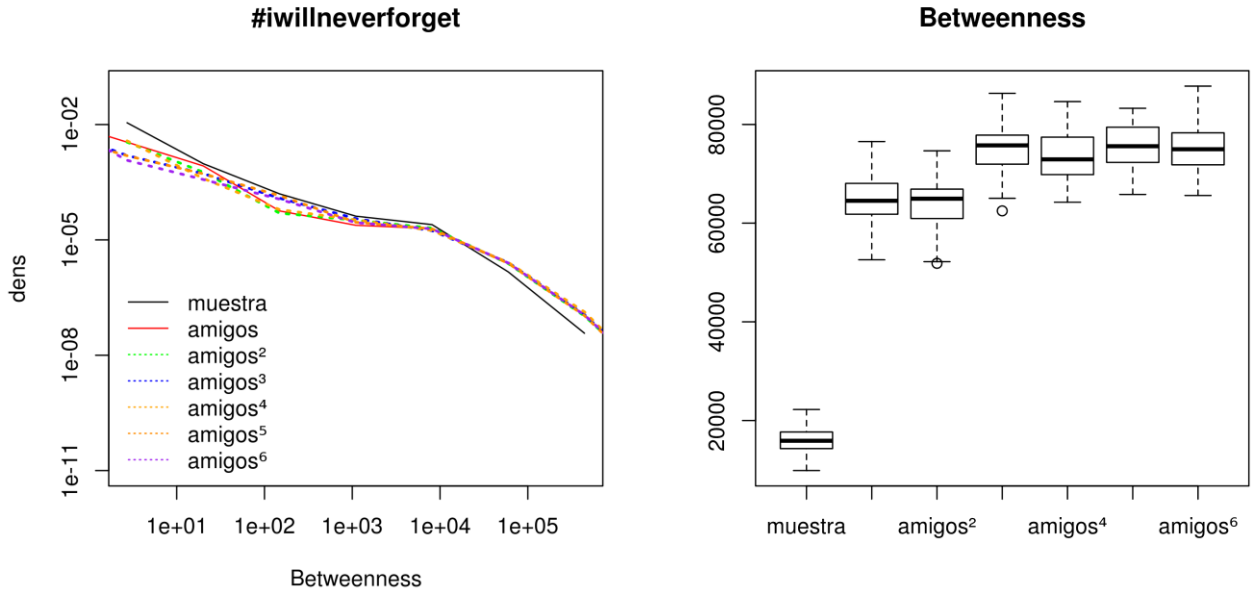


Ilustración 19. Distribución de la betweenness y betweenness medio para cuando se eliminan los usuarios duplicados una vez obtenida la muestra

4.3.1.2 K-Coreness

La medida de centralidad k-coreness tanto para todos los usuarios como para cuando se eliminan duplicados de la muestra de los amigos, se comporta de forma similar. Existe un gran salto entre la *muestra* y los *amigos*, es decir la paradoja de la amistad también funciona en esta medida. Pero ahora los siguientes grupos de amigos sí que penetran más, siendo $amigos^3$ donde se estanca el crecimiento. Los siguientes son muy similares, en cuanto se refiere a centralidad k-coreness.

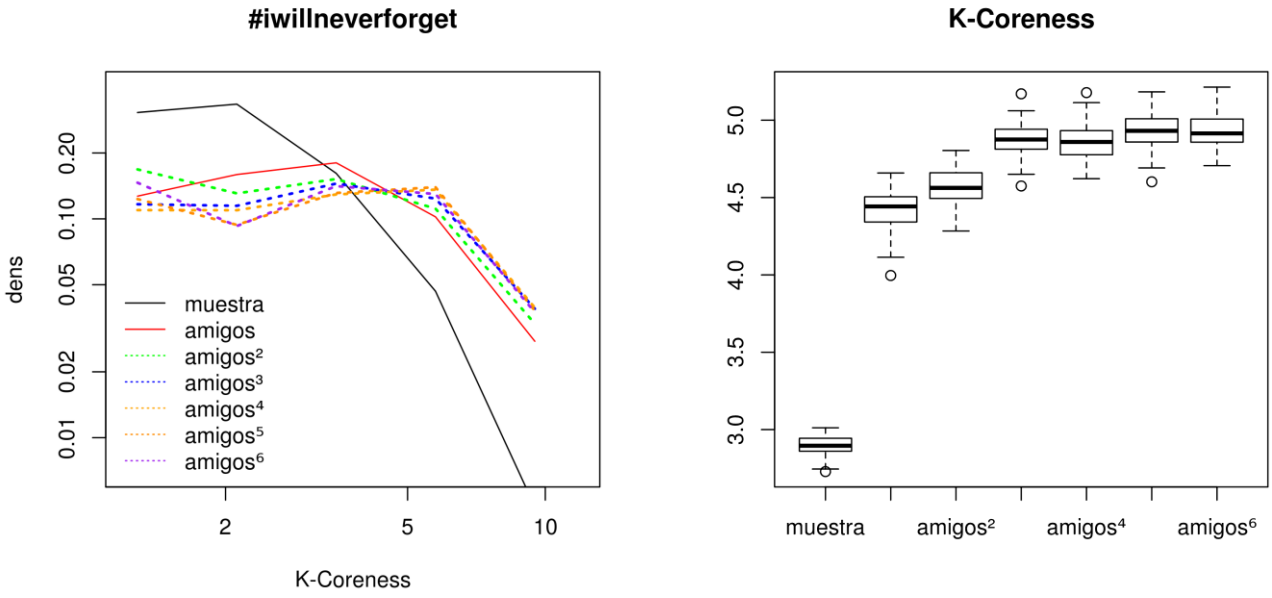


Ilustración 20. Distribución de k-coreness y k-coreness medio para amigos con repetición

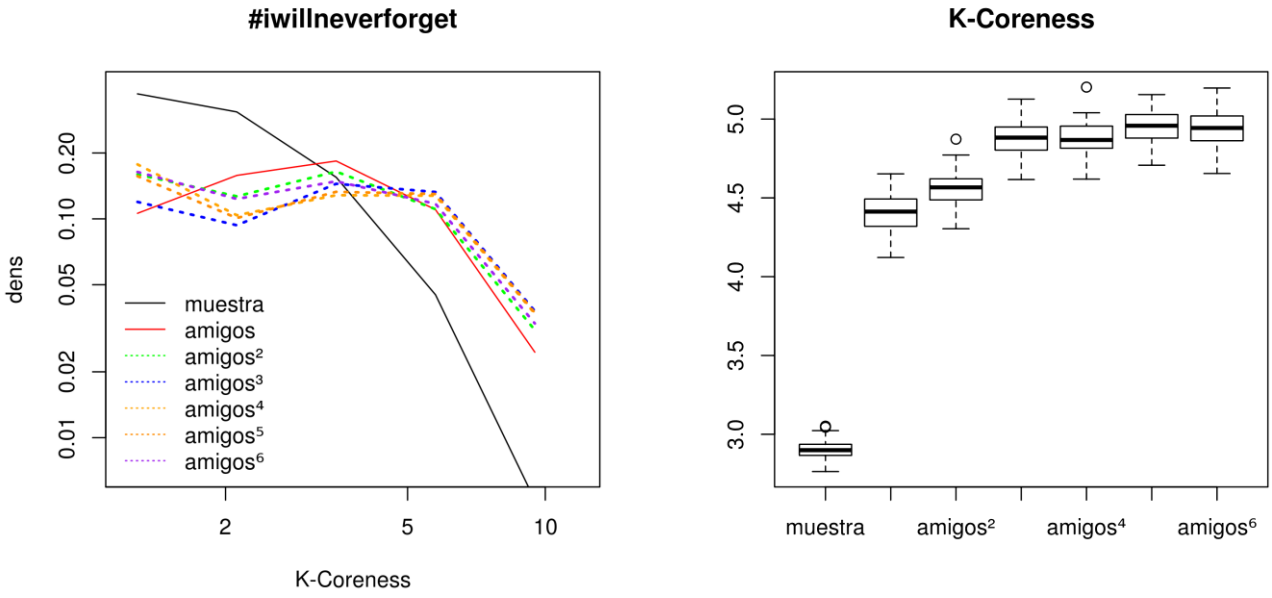


Ilustración 21. Distribución de k-coreness y k-coreness medio para cuando se eliminan los usuarios duplicados una vez obtenida la muestra

4.3.1.3 Closeness

La medida de centralidad closeness es la que más varía. En ella el punto de inflexión lo marca el grupo *amigos*^{2} para las dos estrategias de obtención de amigos. La centralidad dada por closeness parece creciente para el resto de muestras de amigos que se tienen.

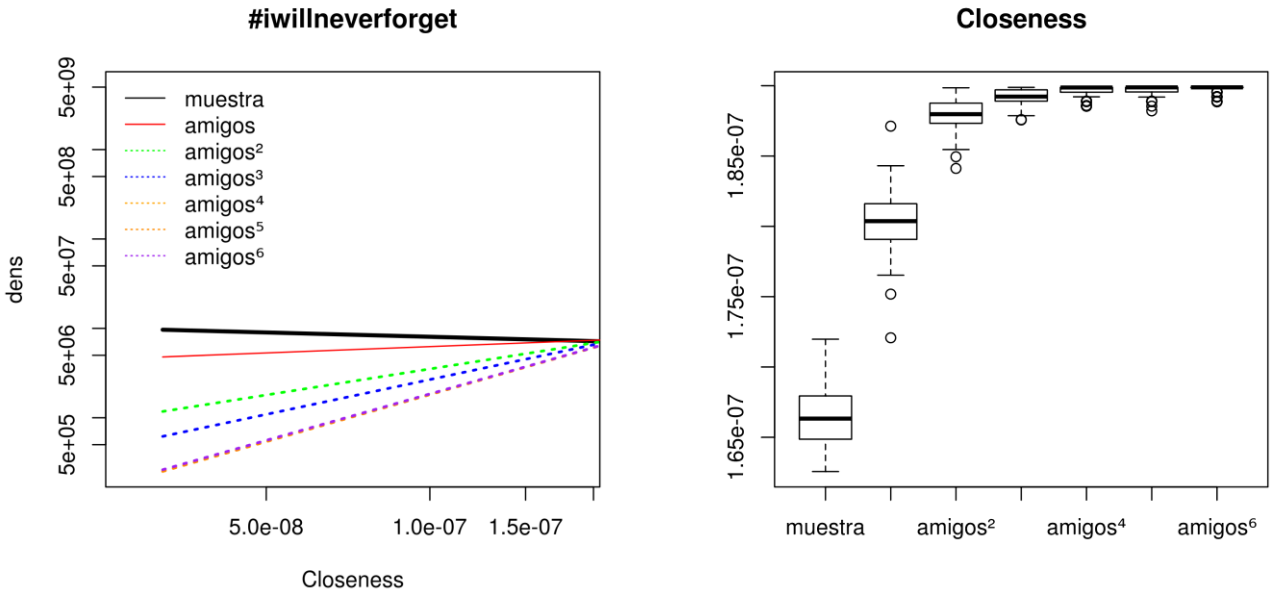


Ilustración 22. Distribución de closeness y closeness medio para amigos con repetición

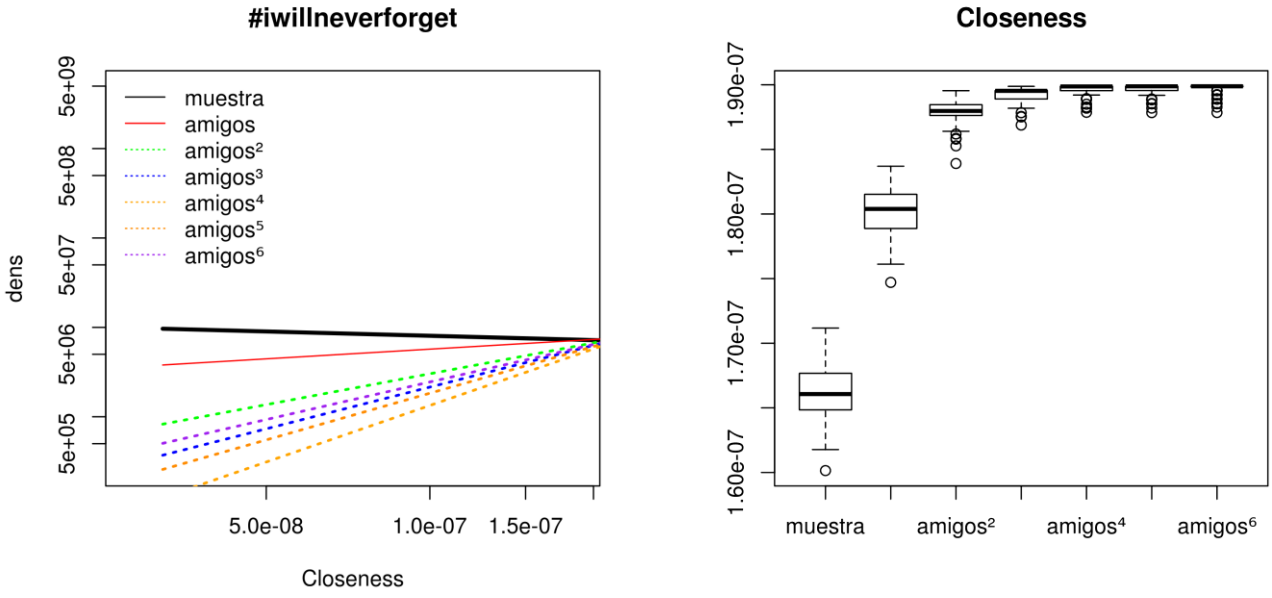


Ilustración 23. Distribución de closeness y closeness medio para cuando se eliminan los usuarios duplicados una vez obtenida la muestra

4.4. Diferencias en la obtención de grupos más centrales

Utilizar la paradoja de la amistad conlleva grandes ventajas. Una de ellas es que se puede encontrar un grupo de individuos más central en la red sin tener que realizar el estudio de la red global.

Se escoge una muestra aleatoria de 50.000 usuarios, de este grupo se obtienen los amigos y de esos amigos se seleccionan dos grupos:

1. 50.000 amigos escogidos aleatoriamente, *rnd_friends.* La media de grado de este grupo fue 836,35**.**
2. Los 50.000 amigos con mayor grado, *top_friends*. El grado medio fue 11.233,5.

Teniendo ya el grado medio de la muestra *top_friends* se procede a buscar qué tamaño de muestra de la red global da como resultado un grado medio similar.

Tabla 3. Grado medio para diferentes tamaños de muestra

Tamaño de la muestra	Grado medio
Muestra(50.000)	**32,62**
500.000	266,50
1.000.000	523,85
Amigos aleatorios	**836,35**
2.500.000	1.226,76
5.000.000	2.070,04
7.500.000	2.951,89
10.000.000	3696,29
20.000.000	6.445,20
Top Amigos	**11.233,5**

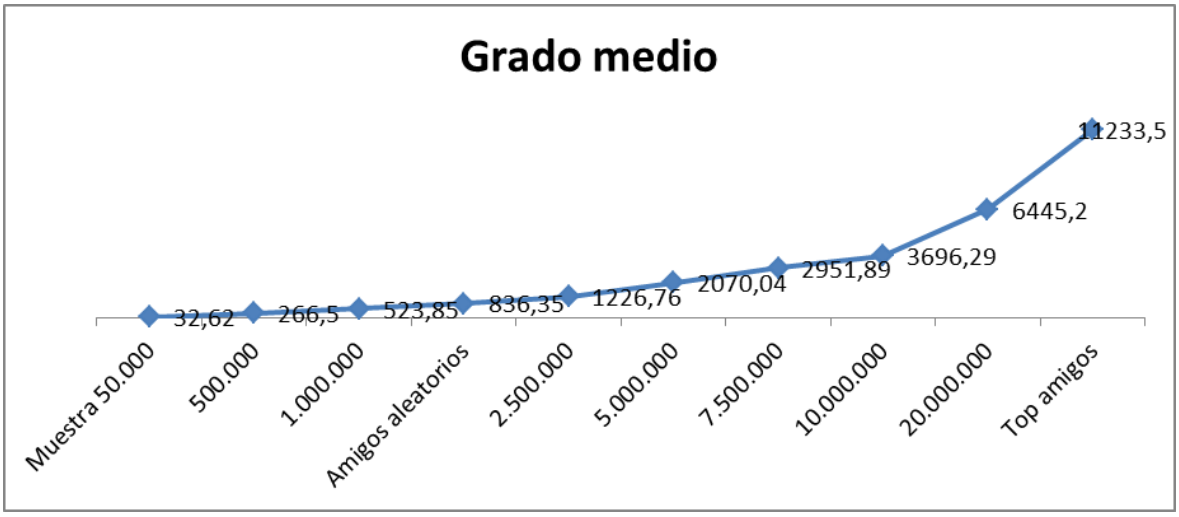


Ilustración 24. Gráfico con la evolución del grado medio para diferentes tamaños de muestra

La muestra que más se acerca al grado medio de *top_friends* es la muestra de tamaño 20 millones y se deduce que con una muestra de aproximadamente 30 millones de usuarios se obtendrá un grado similar.

La principal diferencia en conseguir individuos centrales en una red aplicando un análisis local o uno global está en el tamaño del grupo a explorar. Utilizando la paradoja de la amistad basta un grupo de 50.000 miembros de la red de los cuales obtener sus amigos. En cambio si el análisis es global el tamaño de la muestra tiene que ser de al menos 30.000.000 personas (un 600% más del tamaño inicial al que se le aplicó la teoría de Feld). Esto nos muestra que realizar un análisis global tiene asociado desventajas frente al hacer un estudio local. Las principales son la necesidad de conocer la mayor parte de la red con lo que el coste computacional será mucho mayor.

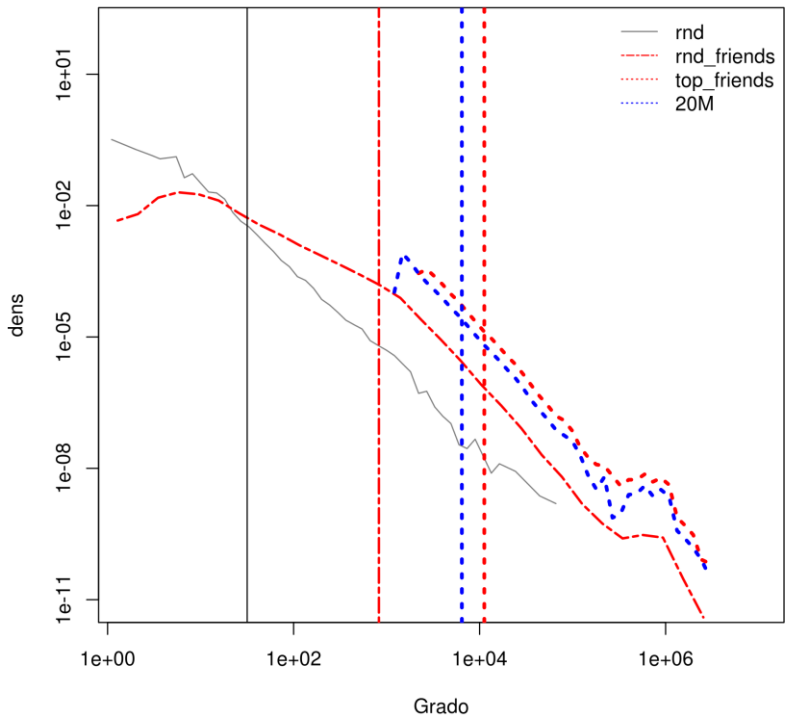


Ilustración 25. Diferencia de grado para muestras de distintos tamaños. Las líneas verticales marcan el grado medio de la muestra inicial (rnd en negro), de una muestra aleatoria de los amigos de la muestra inicial (rnd_friends en rojo), los 50.000 usuarios con mayor grado obtenidos de una muestra de 20 millones (20M en azul) y de los 50.000 amigos de la muestra inicial con mayor grado (top_friends en rojo discontinuo)

5. Conclusiones y trabajo futuro

Una vez terminados todos los análisis la primera conclusión que se recoge, y a mi parecer la más importante y válida, es el coste tan pequeño que requiere hacer un análisis local, en comparación a un análisis global, cogiendo una muestra aleatoria de una red y aplicándole la paradoja de la amistad. Se obtienen usuarios más centrales en la red que proporcionan más información y la obtención de estos usuarios tiene un coste computacional órdenes de magnitud menor.

El grado crece considerablemente con el cálculo de los amigos, pero a partir de ahí no penetra más, sino que muestra una tendencia decreciente perdiendo grado para los siguientes grupos de amigos, concluyendo así que para obtener usuarios con un grado alto bastaría con calcular los amigos de la muestra. Los amigos de tus amigos **no** tienen más amigos que tus amigos.

De las estrategias que se han explorado para obtener los amigos de una muestra, coger todos los amigos con repetidos y después eliminar duplicados es la mejor opción para obtener un grupo con un grado significativamente más alto, a un coste de análisis muy pequeño y sin pervertir las medias con usuarios duplicados.

Como trabajo futuro se podría estudiar si se puede implementar una forma parcial en la obtención de amigos, cogiendo sólo un subconjunto de los amigos de cada nodo en lugar de considerarlos todos y estudiar las repercusiones de esta estrategia. Así se podría reducir el coste de obtener todos los amigos.

Aunque la paradoja no penetra en la distribución de grado para sucesivas iteraciones de la paradoja de la amista, hay algunas medidas de centralidad, como betweennes y k-coreness, en las que existe un salto en la tercera iteración donde se consigue un incremento en la centralidadrelativo, aunque después de este grupo la tendencia general suele ser decreciente. Con el closenes, cabe destacar que se ha encontrado un comportamiento diferente, independiente además de la estrategias en la selección de los grupos de amigos seguida. En este se observa un crecimiento asintótico en el que cada iteractión mejora a la anterior.

Como parte del trabajo futuro, aparte de hacer un estudio más en profundidad en las diferentes formas de seleccionar los amigos de un grupo dado que ayudará a reducir costes, también sería interesante ampliar este estudio a otras redes como pueden ser redes de *retweets* (mensaje de un usuario que es compartido por otros usuarios en su Twitter) o en redes de menciones entre usuarios, así como el estudio de otras medidas no relacionadas tan directamente con la centralidad como son el uso de hashtags, cantidad de tweets geolocalizados o cantidad de viajes.

Glosario

Amigos: amigos de la muestra inicial

Amigos^{2}: amigos de los amigos de la muestra inicial

Amigos^{3}: amigos de los amigos de los amigos de la muestra inicial

Amigos^{4}: amigos de los amigos de los amigos de los amigos de la muestra inicial

Amigos^{5}: amigos de los amigos de los amigos de los amigos de los amigos de la muestra inicial

Amigos^{6}: amigos de los amigos de los amigos de los amigos de los amigos de los amigos de la muestra inicial

Bibliografía

[1] L. Sun, K. W. Axhausen, D.-H. Lee y M. Cebrian, «Efficient detection of contagious outbreaks in massive metropolitan encounter networks,» *Scientific Reports*, 2014.

[2] J. Podesta, P. Pritzker, E. J. Moniz, J. Holdren y J. Zients, «Big Data: Seizing Opportunities, Preserving Values,» Mayo 2014. [En línea]. Available: http://images.politico.com/global/2014/05/01/big_data_privacy_report_may_1_2014.html.

[3] N. O. Hodas, F. Kooti y K. Lerman, «Friendship Paradox Redux: Your Friends Are More Interesting Than You,» *ICWSM*, vol. 13, pp. 8-10, 2013.

[4] D. S. White, «Social Media Growth 2006 to 2012,» 09 02 2013. [En línea]. Available: http://dstevenwhite.com/2013/02/09/social-media-growth-2006-to-2012/.

[5] F. A. Sarría, «Introducción a AWK,» [En línea]. Available: http://ocw.um.es/gat/contenidos/ldaniel/ipu_docs/la_shell/awk.pdf.

[6] «Introducción a R,» [En línea]. Available: http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf.

[7] F. Kooti, N. O. Hodas y K. Lerman, «Network Weirdness: Exploring the Origins of Network Paradoxes,» *arXiv preprint arXiv:1403.7242*, 2014.

[8] Y. Kryvasheyeu, H. Chen, E. Moro, P. Van Hentenryck y M. Cebrian, «Performance of Social Network Sensors During Hurricane Sandy,» *arXiv preprint arXiv:1402.2482*, 2014.

[9] N. A. Christakis y J. H. Fowler, «Social Network Sensors for Early Detection of Contagious Outbreaks,» *PloS one*, vol. 5, nº 9, 2010.

[10] «Twitter,» Marzo 2006. [En línea]. Available: http://www.twitter.com.

[11] M. Garcia-Herranz, E. Moro Egido, M. Cebrian, N. A. Christakis y J. H. Fowler, «Using Friends as Sensors to Detect Global-Scale,» *PloS one*, vol. 9, nº 4, 2014.

[12] S. L. Feld, «Why your friends have more friends than you do,» *American Journal of Sociology*, 1991.

[13] H. Kwak, L. Changhyun, P. Hosung y S. Moon, «What is Twitter, a social network or a news media?,» de *Proceedings of the 19th international conference on Worl Wide Web*, 2010.